【課題番号】課題１

【課題名】実データとオントロジーを自動的に対応付けるための支援ツール開発

１）課題とその背景

バイオサイエンスデータベースセンター(NBDC)の統合化推進プログラムでは、セマンティック・ウェブ技術を基盤としたデータの統合化が進んでいる。様々な生物学の概念が OWL (Web Ontology Language)によるオントロジーを用いて記述されるとともに、それを用いて実データが RDF (Resource Description Framework) 化されている。データを統合化する際、実データを既存オントロジーまたは研究者が構築した新規オントロジーへ対応付ける（マッピングする）ことは、異なったデータベース由来のデータを概念の共通性から統合化する上で不可欠な作業である。基本的に、この作業は実データおよびオントロジーに関わる専門家が手作業で行っているのが現状である。一例として微生物の統合データベースである MicrobeDB.jp では、微生物の生息環境に関する様々なメタデータと様々なオントロジーとのマッピングを手作業で行っている。しかしながら、オントロジーの規模が大きくなったり、マッピング対象のデータの数が多くなるにつれて非常に手間と時間がかかるようになり、マッピングを手作業で行うことは現実的ではなくなる。既に MicrobeDB.jp においても手作業では困難な規模に達してきており、定期的なデータベースのアップデートを行う上で、大きな律速要因になりつつある。この問題は、MicrobeDB.jp に限らず、セマンティック・ウェブを用いてデータの統合化を行う際に共通して生じうる。そこで、ユーザが入力した OWL 形式のオントロジーファイルとマッピング対象の実データの RDF ファイルから、ある程度自動でオントロジーと実データとのマッピングを行う、オントロジーマッピング支援ツールの開発を行う。

２）課題の解決方法の概要

NCBO Annotator や EBI ZOOMA、OpenRefine など、オントロジーと実データとのマッピング支援ツールは既にいくつか存在するが、未だ標準となるようなツールは存在しない。その理由として、既存のマッピングは概念の共通性を基にしているが、文字列としては全く異なる語彙が同義語または類義語である場合も多いほか、実データが単語ではなく文章の場合もあるなど、単語の文字列にもとづくマッピングだけではツールとしてあまり有用なものにならない点などが挙げられる。

前提として、オントロジーと実データは通常英語で記述されているため、英語に対する既存の種々の自然言語処理の技法およびコーパスを用いて前処理を行う必要がある。特に、マッピングしたい実データは単語ではなく文章である場合が多いため、マッピング前に文章の形態素解析を行うことは必須となる。これらの自然言語処理の技術を実装・適用する

ことに加えて、ユーザが以前に手動で行ったオントロジーと実データとのマッピング結果が存在する場合には、そのファイルを参考情報として読み込んで、同様の組み合わせが実データとオントロジーとの間にも見つかった場合には、例え文字列としては別物だったとしても、マッピング候補として表示するなどの工夫が必要になる。また、オントロジーは概念間の階層構造や同義語などの記述を含んでいるため、それらの構造を考慮した上での自動マッピングを行う必要がある。

３）課題の解決方法の概略図